DEEP ENCODED LINGUISTIC AND ACOUSTIC CUES FOR ATTENTION BASED END TO END SPEECH EMOTION RECOGNITION

ICASSP - 2020
Paper code - 5598

Authors
- Swapnil Bhosale
- Rupayan Chakraborty
- Sunil Kumar Kopparapu

TCS Research and Innovation - Mumbai, India

Introduction

- Speech Emotion Recognition (SER) has several applications
  - man-machine interactions
  - human health assistance
  - call center analytics etc.

Introduction

- Speech Emotion Recognition (SER) has several applications
  - man-machine interactions
  - human health assistance
  - call center analytics etc.

- Developments in deep learning especially in terms of,
  - data augmentation
  - better feature extractors
  - cross-domain knowledge transfer

  have significantly impacted SER.

Introduction

- Speech Emotion Recognition (SER) has several applications
  - man-machine interactions
  - human health assistance
  - call center analytics etc.

- Developments in deep learning especially in terms of,
  - data augmentation
  - better feature extractors
  - cross-domain knowledge transfer

  have significantly impacted SER.

- Can be further improved by exploiting,
  - ➢ Acoustic information : Spectrograms from raw audio and glottal source signals
  - ➢ Linguistic information : Text, Phoneme sequences, intermediate DNN representations

Related Work

Two directions :

- Use complex hand-crafted features (ex: OpenSMILE feature set)
- Deep modelling with conventional raw audio spectrograms

Related Work

Two directions :

- Use complex hand-crafted features (ex: OpenSMILE feature set)
- **Deep modelling with conventional raw audio spectrograms (End-to-End)**

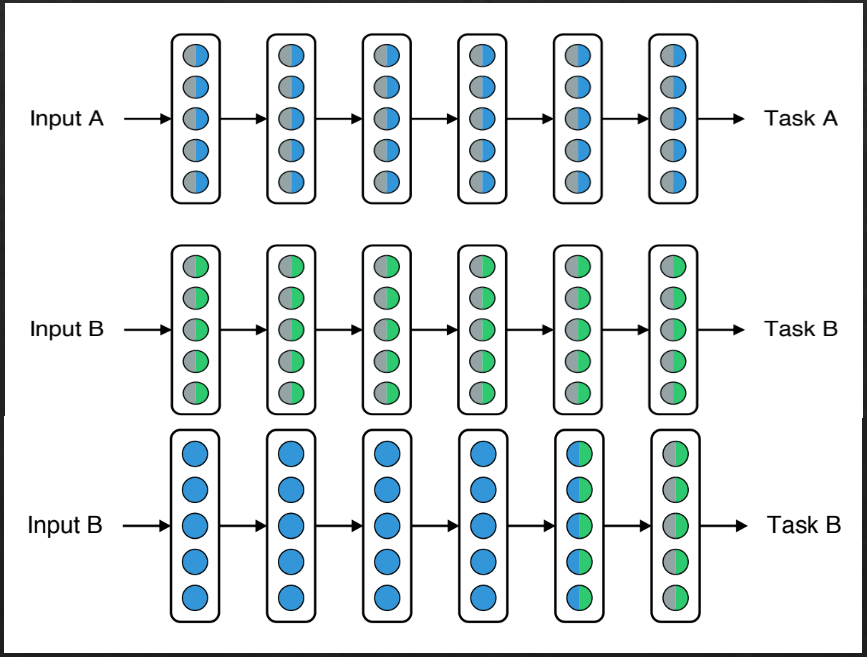
Related Work

Two directions :

- Use complex hand-crafted features (ex: OpenSMILE feature set)
- **Deep modelling with conventional raw audio spectrograms (End-to-End)**

Transferring knowledge within tasks/datasets[1]

- In Deep networks,
  - initial layers ➔ low-level features
  - final layers ➔ high-level features
- Transfer learning ➔ share knowledge across datasets and tasks.



[1] https:/ /haythamfayek.com/assets/talks/Fayek_neurips18.pdf

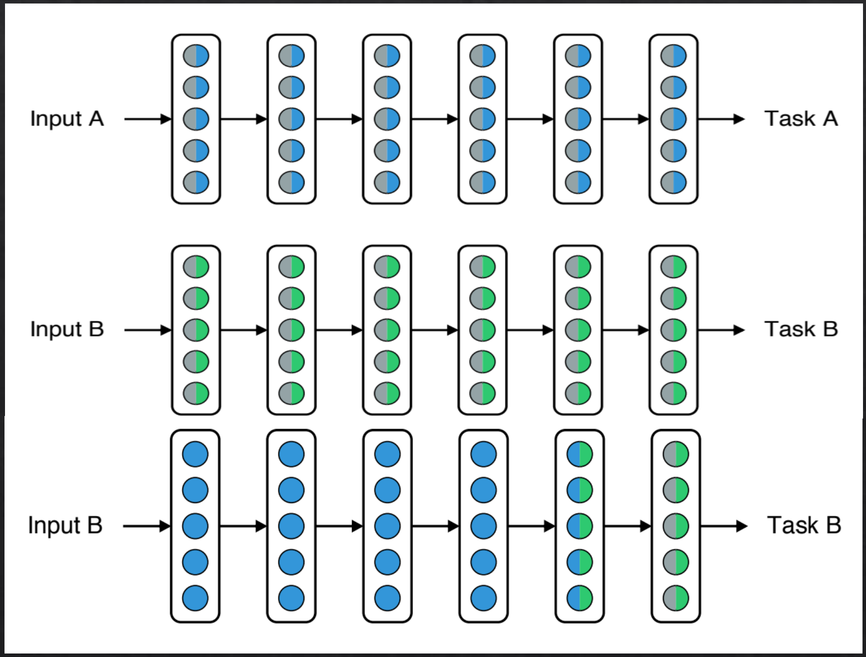
Related Work

Two directions :

- ➢ Use complex hand-crafted features (ex: OpenSMILE feature set)
- ➢ **Deep modelling with conventional raw audio spectrograms (End-to-End)**

Transferring knowledge within tasks/datasets[1]

- In Deep networks,
  - initial layers ➔ low-level features
  - final layers ➔ high-level features
- Transfer learning ➔ share knowledge across datasets and tasks.
- Objective : Maximum knowledge transfer, minimum dependency on parent task/dataset.



[1] https:/ /haythamfayek.com/assets/talks/Fayek_neurips18.pdf

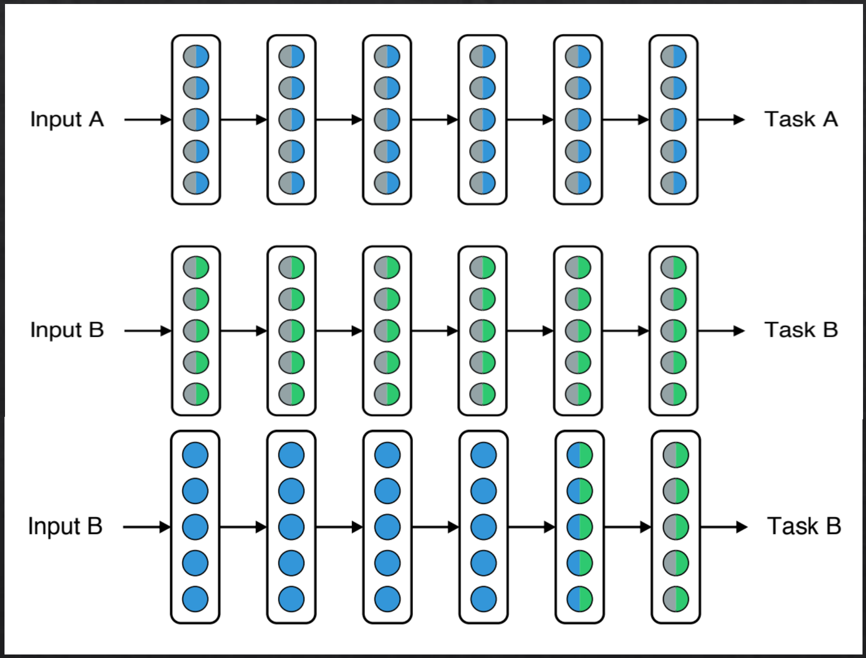
Related Work

Two directions :

- ➢ Use complex hand-crafted features (ex: OpenSMILE feature set)
- ➢ **Deep modelling with conventional raw audio spectrograms (End-to-End)**

Transferring knowledge within tasks/datasets[1]

- In Deep networks,
  - initial layers ➔ low-level features
  - final layers ➔ high-level features
- Transfer learning ➔ share knowledge across datasets and tasks.
- Objective : Maximum knowledge transfer, minimum dependency on parent task/dataset.

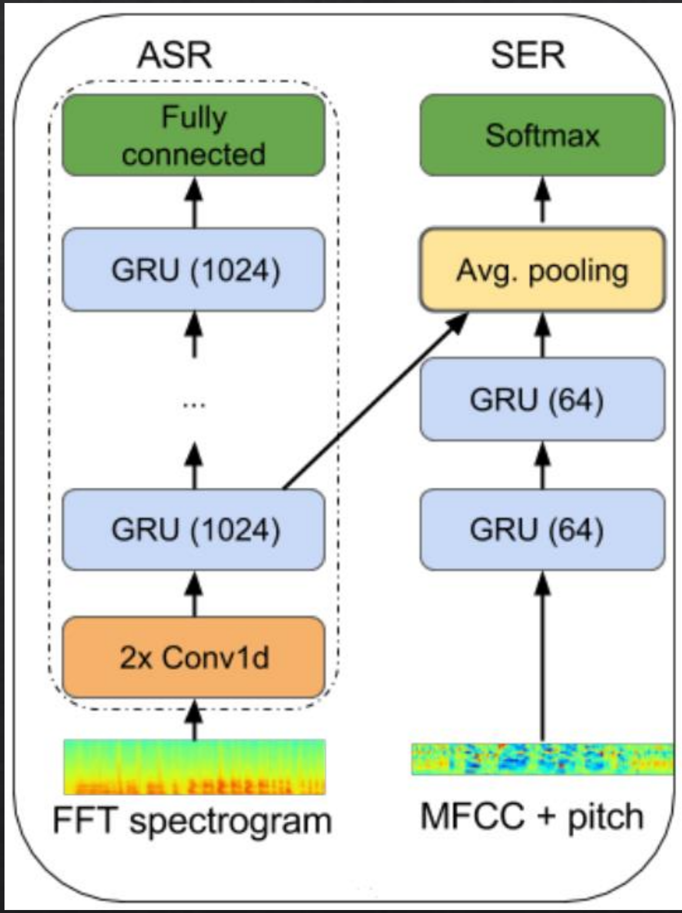


"Low-level features are more generic and easier to transfer from one task to another"
Could there be exceptions?

Related Work

Jointly learning supplementary tasks [2]

- Uncertainty about most relevant and robust features/layers
- Progressive network : training ASR and SER tasks jointly
- ASR representations show improved performance mainly due to the robustness to speaker and condition variations.

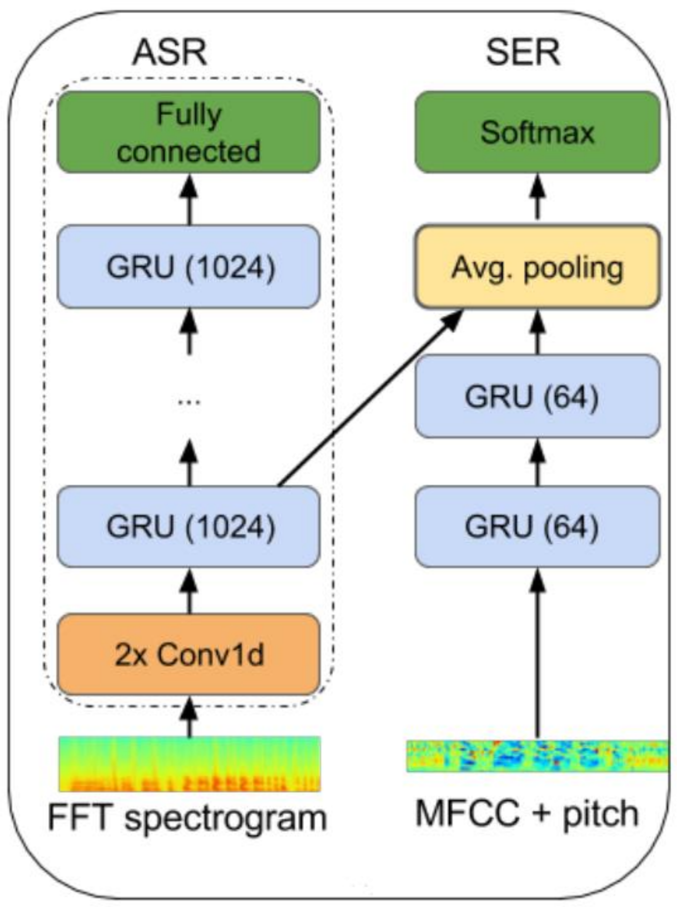


[2] https://www.aclweb.org/anthology/I17-1043.pdf

Related Work

Jointly learning supplementary tasks [2]

- Uncertainty about most relevant and robust features/layers
- Progressive network : training ASR and SER tasks jointly
- ASR representations show improved performance mainly due to the robustness to speaker and condition variations.



Key Takeaways from related work:

- Influence of linguistic knowledge in spoken utterances for SER task still remains unexplored.
- Selection of intermediate ASR layers needs to be studied thoroughly.

[2] https://www.aclweb.org/anthology/I17-1043.pdf

Proposed System

Representative features

Acoustic features : Mel-spectrogram

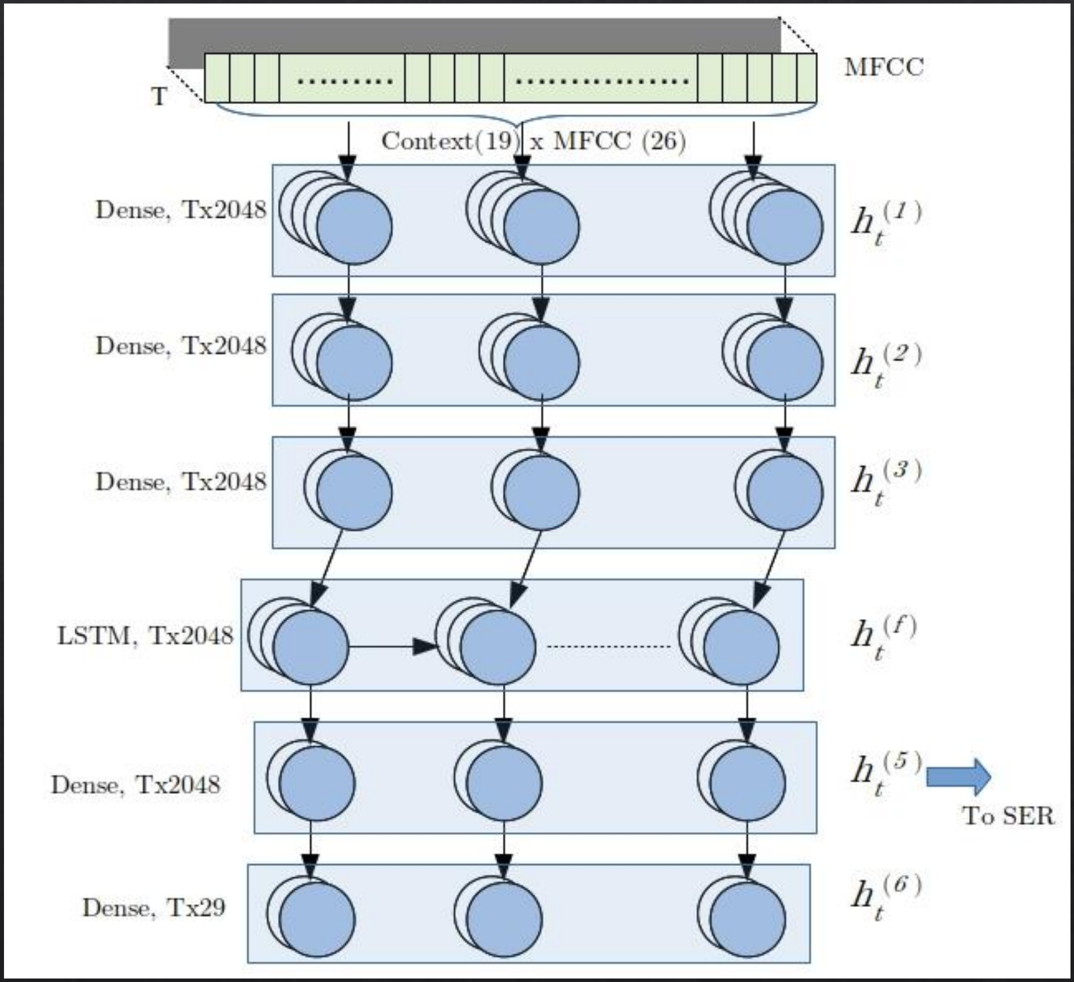
- Sampling rate = 16 kHz
- Frame duration = 25 msec
- Length of FFT window = 2048
- Hop length = 400 samples
- Number of bins on mel-scale = 128

Concatenate Δ and Δ- Δ for the mel-spectrogram.

Representative features

Deep encoded Linguistic features : DeepSpeech ASR [3]

- Note : Layers closer to output capture the linguistic content of speech while the layers close to input capture the acoustic content.[4]



DeepSpeech–1 architecture

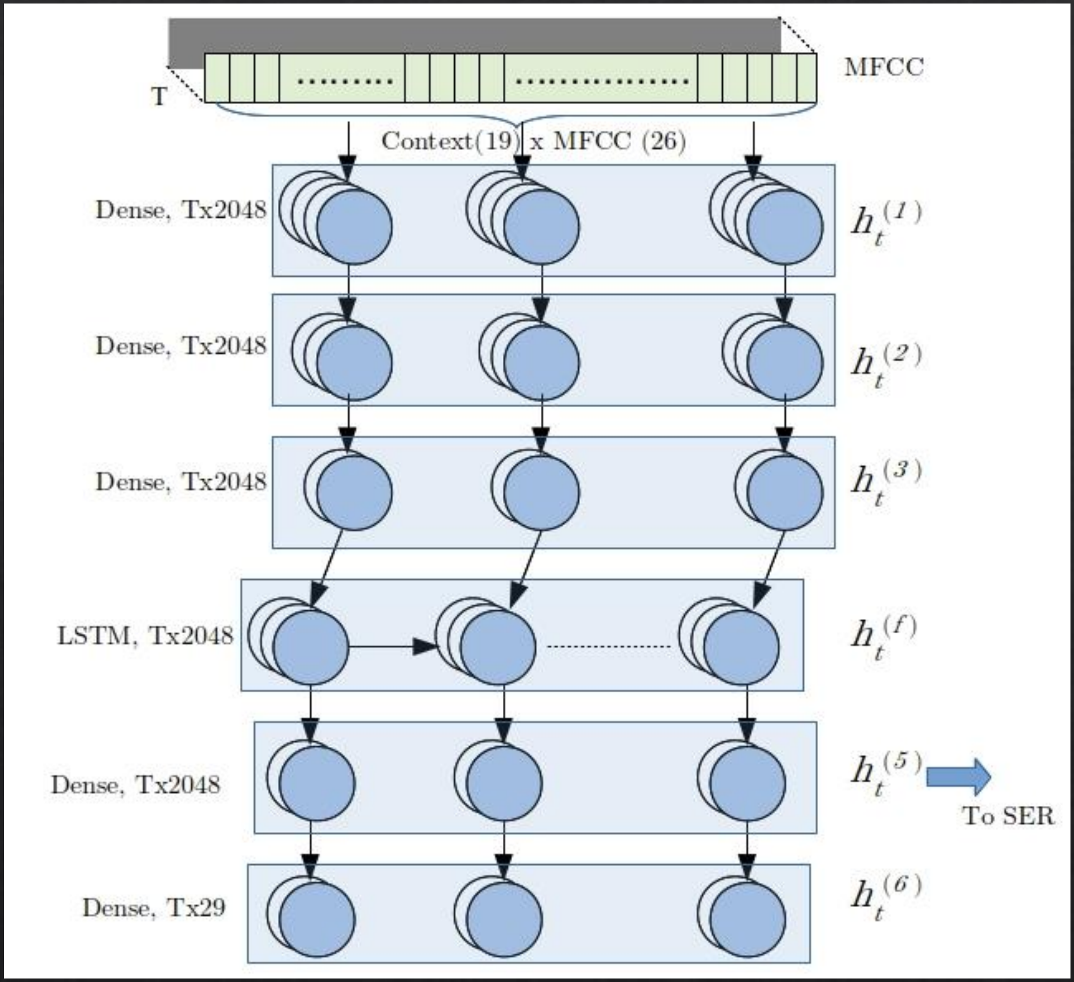
[3] Mozilla, “DeepSpeech–0.4.0,”https://github.com/mozilla/DeepSpeech/releases, January 2019
[4] Swapnil B, Imran S, Sunil K, “End-to-End spoken language understanding: Bootstrapping in low resource scenarios,” Interspeech 2019.

Representative features

Deep encoded Linguistic features : DeepSpeech ASR [3]

- Note : Layers closer to output capture the linguistic content of speech while the layers close to input capture the acoustic content. [4]

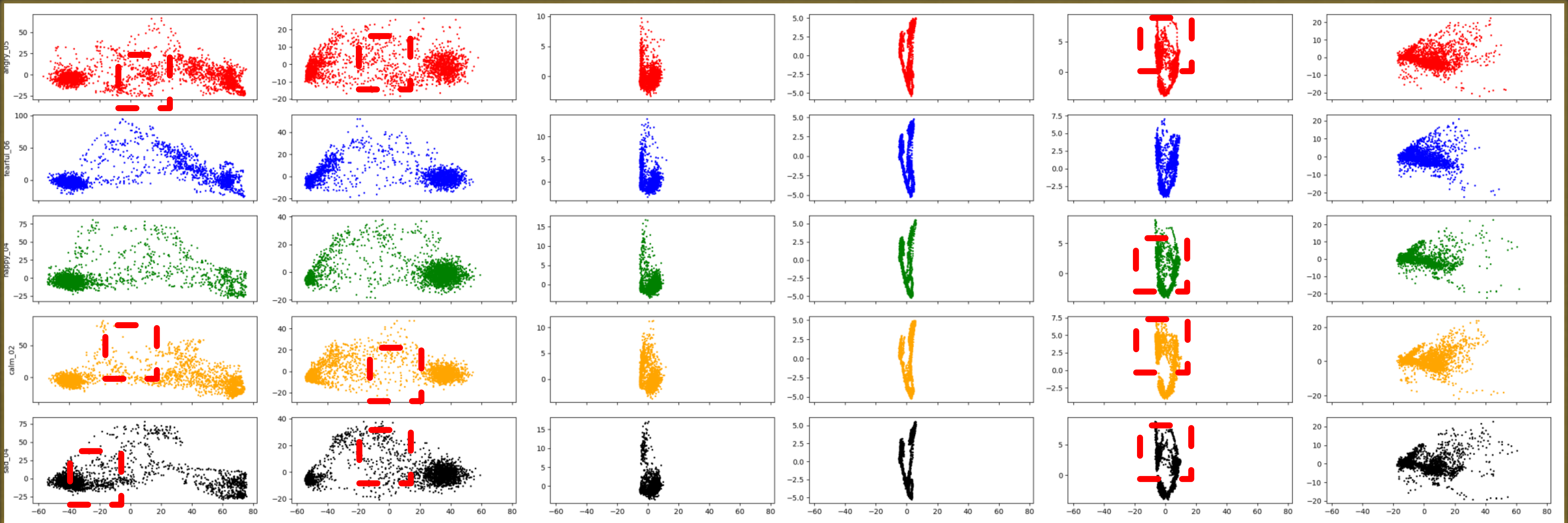
Can we get linguistic context of embedded emotion in the spoken utterance ?



DeepSpeech–1 architecture

[3] Mozilla, "DeepSpeech–0.4.0,"https://github.com/mozilla/DeepSpeech/releases, January 2019
[4] Swapnil B, Imran S, Sunil K, "End-to-End spoken language understanding: Bootstrapping in low resource scenarios," Interspeech 2019.

Representative features



Visualization of activations from different layers of DeepSpeech model, for the same utterance spoken in different emotions. Columns represent the 6 layers and rows represent emotions. *anger*; *fearful*; *happy*; *calm*; *sad*

- 1st , 2nd and 5th layers show least correlation across the rows (emotions).
- Lesser correlation in 1st and 2nd layer is due to variations in speaker, gender etc.[4]
- We use the output from the 5th layer for getting the linguistic context for the SER task.

Proposed architecture

Encoder :

- 2 layers of 1-D convolutions.
  - ➢ Helps to learn temporal context between adjacent frames.

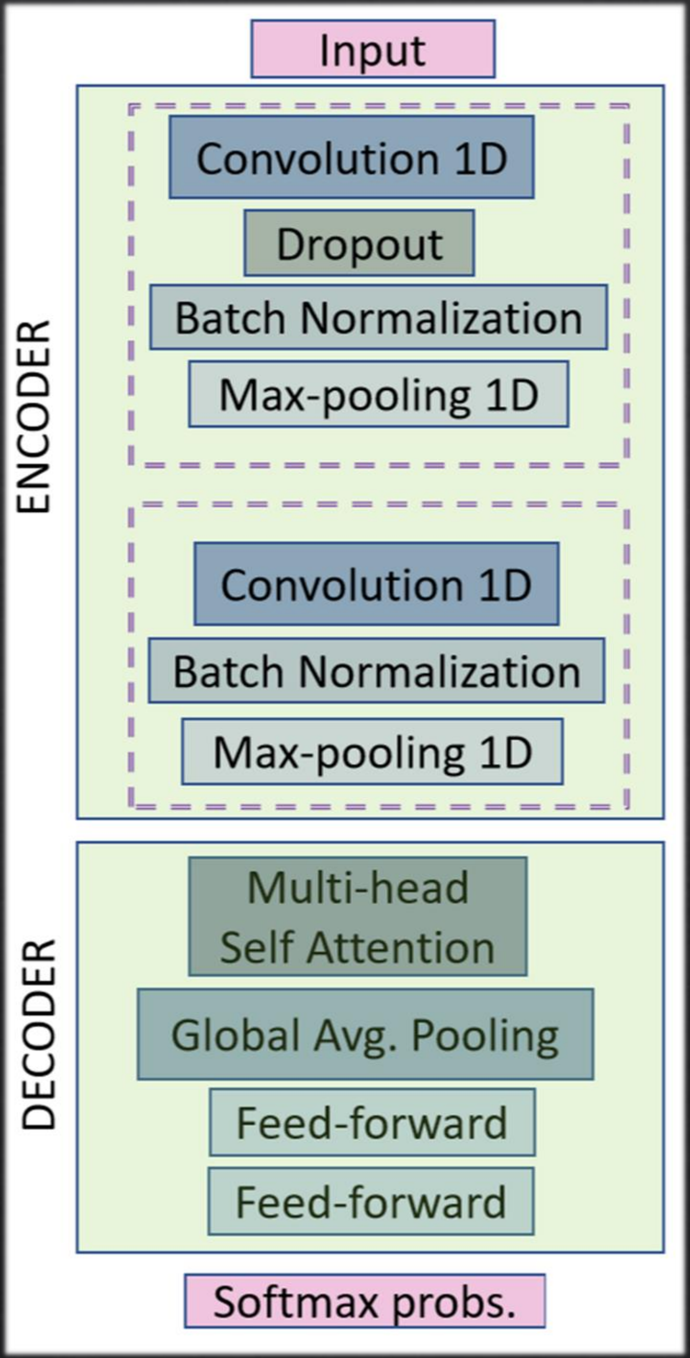
- 1-D convolution layer
  ➔ Batch normalization layer
  ➔ ReLU activation

Decoder :

- Multi-head self attention layer
  ➔ Average pooling
  ➔ 2 feedforward dense layers.

Output :

Softmax distribution over individual emotions.



Proposed Encoder – Decoder model architecture

Multi-head Self Attention

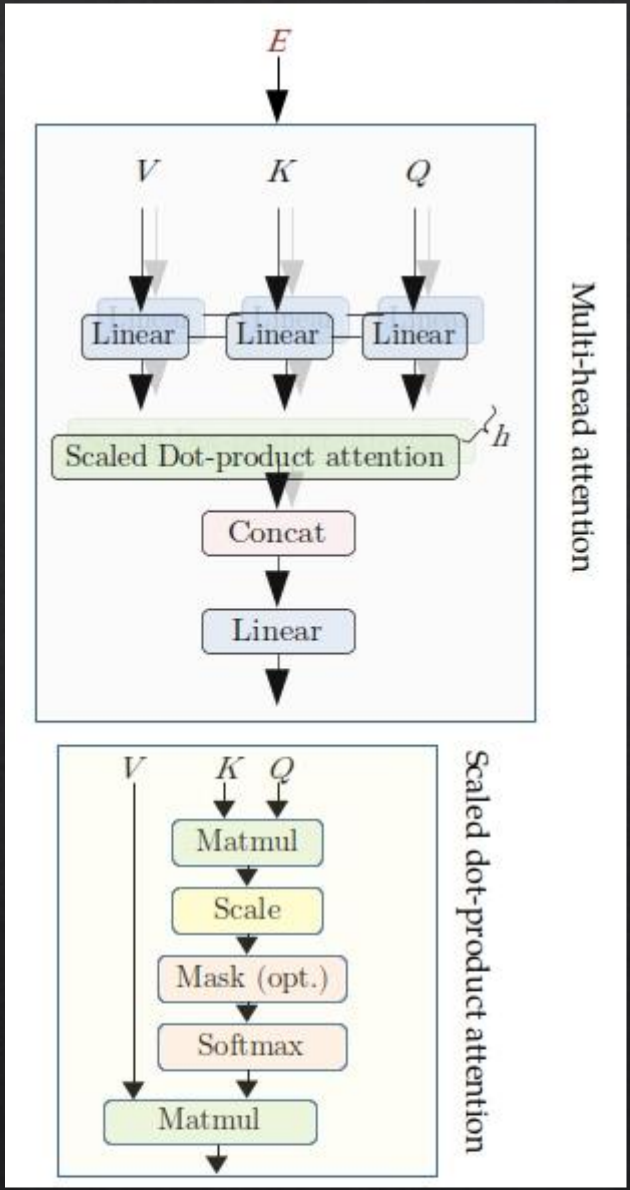
- Let E be the output of the encoder block
- W_i are trainable weight matrices
- d_i is the dimension
- A_i : Attention weight of a single head
- A_{MH} : Final multi-head self attention
- h : total number of heads

$$A_i = softmax\left(\frac{Q_i K_i^T}{\sqrt{d_K}}\right) V_i \ \forall i \in \{1,2, \ldots \text{h}\}$$

$$Q = EW_Q, \quad K = EW_K, \quad V = EW_V$$

$$A_{MH} = (A_1 \| A_2 \| \ldots \| A_h) W_E$$

$$\text{Context, } C = E + A_{MH}$$



Multi-head Self Attention

- Let E be the output of the encoder block
- W_i are trainable weight matrices
- d_i is the dimension
- A_i : Attention weight of a single head
- A_{MH} : Final multi-head self attention
- h : total number of heads

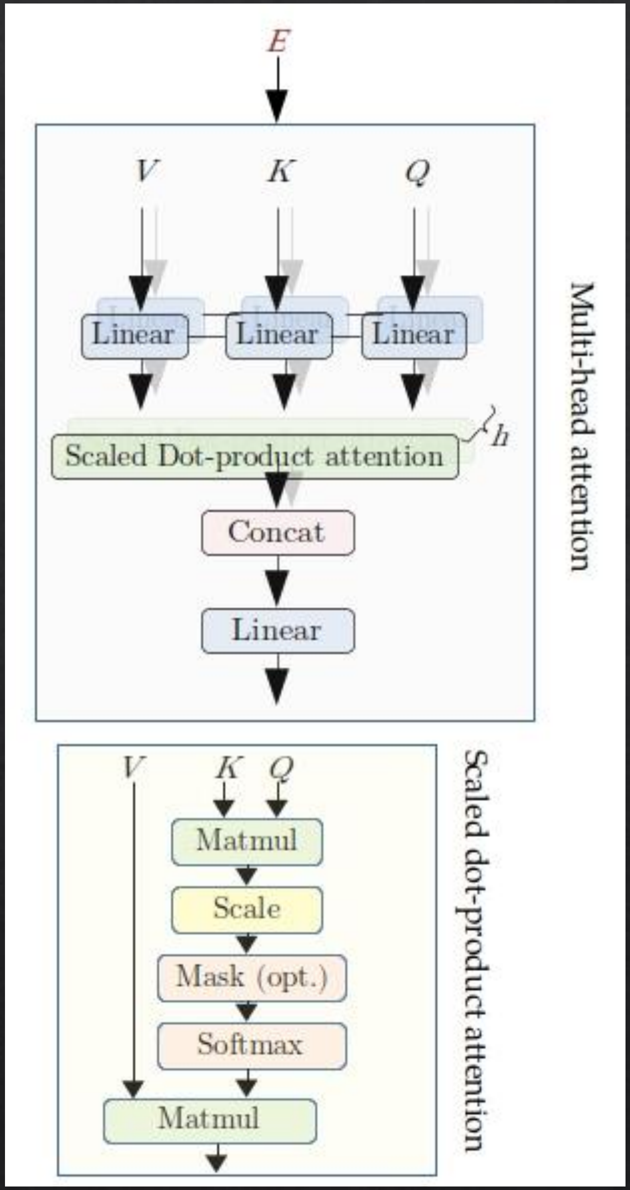
$$A_i = softmax\left(\frac{Q_i K_i^T}{\sqrt{d_k}}\right) V_i \ \forall i \in \{1,2, \ldots \text{h}\}$$

$$Q = EW_Q, \quad K = EW_K, \quad V = EW_V$$

$$A_{MH} = (A_1 \| A_2 \| \ldots \| A_h) W_E$$

$$\text{Context, } C = E + A_{MH}$$

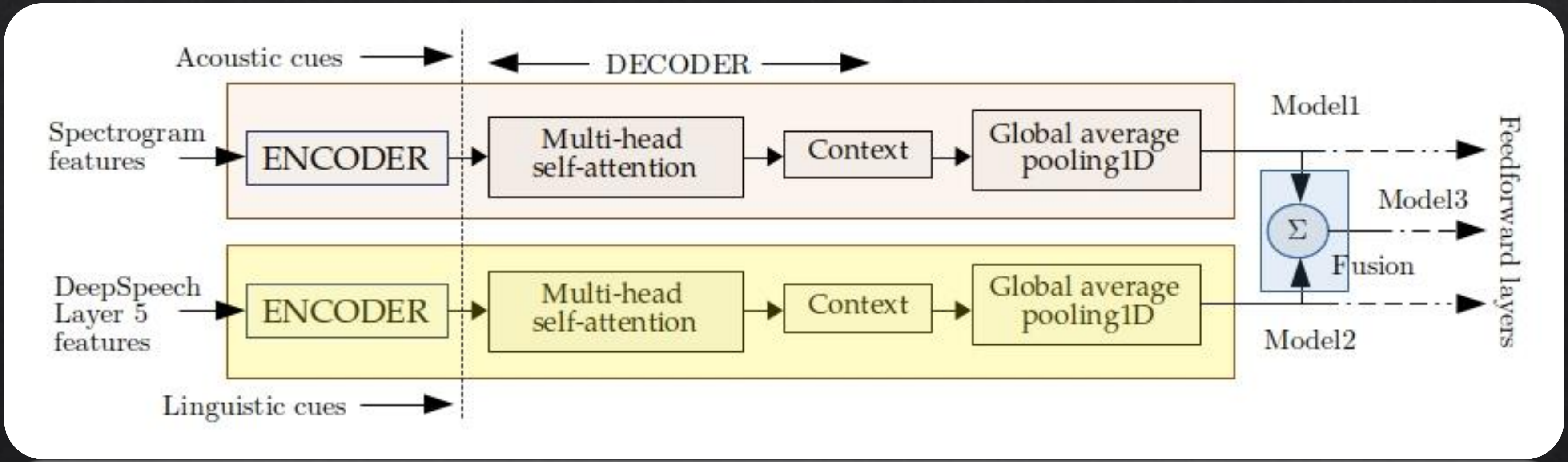
Var.	Dim.
E_i	[t X d_E]
Q_i	[t X d_Q]
K_i	[t X d_K]
$Q_i K_i^T$	[t X t]
V_i	[t X d_V]
$(Q_i K_i^T) V_i$	[t X d_V]



Experiments

- Dataset : IEMOCAP[5]
  - Recording setups : 2 { Improvised speech, scripted play }
  - Categorical Emotion classes : 4 { anger, happiness, neutral, sadness }

- Model configurations :



Model configurations : Model-1, Model-2, Model-3

[5] Carlos Busso, IEMOCAP: Interactive Emotional Dyadic Motion Capture database," Language resources and evaluation, 2008.

Results

Experiments with improvised recordings

Model (Input features)	Weighted Acc., WA	Unweighted Acc., UA
Yenigalla et al.,2018 [6] (only spectrogram)	71.3	61.6
Satt et al., 2017 [7]	68.8	59.4
Lee et al., 2015 [8]	63.8	62.85
Model - 1 (acoustic)	72.08	58.53
Model - 1 (downsampling + ensembling)	70.05	**63.27**
Model - 2 (linguistic)	69.56	54.62
Model - 3 (fusion)	**72.34**	58.31

Observation :

- Improvement using only Acoustic features ✓
- Improvement using Linguistic features (or +Acoustic features) ✗

[6] Promod Yenigalla, "Speech emotion recognition using spectrogram & phoneme embedding,". Interspeech 2018
[7] Aharon Satt, "Efficient emotion recognition from speech using deep learning on spectrograms ,". Interspeech 2017
[8] Jinkyu Lee, "High-level feature representation using recurrent neural network for speech emotion recognition,". Interspeech 2015

Results

Experiments with improvised recordings

Model (Input features)	Weighted Acc., WA	Unweighted Acc., UA
Yenigalla et al.,2018 [6] (only spectrogram)	71.3	61.6
Satt et al., 2017 [7]	68.8	59.4
Lee et al., 2015 [8]	63.8	62.85
Model - 1 (acoustic)	72.08	58.53
Model - 1 (downsampling + ensembling)	70.05	**63.27**
Model - 2 (linguistic)	69.56	54.62
Model - 3 (fusion)	**72.34**	58.31

Observation :

- Improvement using only Acoustic features ✓
- Improvement using Linguistic features (or +Acoustic features) ✗

Reasoning : Improvised recordings carry less linguistic correlations and capture emotion representative characteristics mostly in acoustic space.

[6] Promod Yenigalla, “Speech emotion recognition using spectrogram & phoneme embedding,”. Interspeech 2018
[7] Aharon Satt, “Efficient emotion recognition from speech using deep learning on spectrograms ,”. Interspeech 2017
[8] Jinkyu Lee, “High-level feature representation using recurrent neural network for speech emotion recognition,”. Interspeech 2015

Results

Experiments with improvised recordings

Model (Input features)	**Weighted Acc., WA**	**Unweighted Acc., UA**
Yenigalla et al.,2018 [6] (only spectrogram)	71.3	61.6
Satt et al., 2017 [7]	68.8	59.4
Lee et al., 2015 [8]	63.8	62.85
Model - 1 (acoustic)	72.08	58.53
Model - 1 (downsampling + ensembling)	70.05	**63.27**
Model - 2 (linguistic)	69.56	54.62
Model - 3 (fusion)	**72.34**	58.31

Observation :

- Improvement using only Acoustic features ✓
- Improvement using Linguistic features (or +Acoustic features) ×

Reasoning : Improvised recordings carry less linguistic correlations and capture emotion representative characteristics mostly in acoustic space.

What if there is linguistic context embedded within the samples?

Results

Experiments with scripted recordings

Model (Input features)	Weighted Accuracy, WA	Unweighted Accuracy, UA
Model - 1 (acoustic)	63.04	52.73
Model - 2 (linguistic)	**68.56**	**60.37**
Model - 3 (fusion)	67.12	59.02

<u>Observation</u> :

- Improvement using only Linguistic features ✓
- Improvement using Acoustic features (or + Linguistic features) ✗

Results

Experiments with scripted recordings

Model (Input features)	Weighted Accuracy, WA	Unweighted Accuracy, UA
Model - 1 (acoustic)	63.04	52.73
Model - 2 (linguistic)	**68.56**	**60.37**
Model - 3 (fusion)	67.12	59.02

<u>Observation</u> :

- Improvement using only Linguistic features ✓
- Improvement using Acoustic features (or + Linguistic features) ✗

<u>Reasoning</u> : Utterances in different sessions but belonging same emotions have similar linguistic content.

➢ 7.64% improvement compared to "only acoustic features" as input.

Results

Experiments with scripted recordings

Model (Input features)	Weighted Accuracy, WA	Unweighted Accuracy, UA
Model - 1 (acoustic)	63.04	52.73
Model - 2 (linguistic)	**68.56**	**60.37**
Model - 3 (fusion)	67.12	59.02

Observation :

- Improvement using only Linguistic features ✓
- Improvement using Acoustic features (or + Linguistic features) ✗

Reasoning : Utterances in different sessions but belonging same emotions have similar linguistic content.

➢ 7.64% improvement compared to “only acoustic features” as input.

What if the data itself has a combination of both scripted and improvised speech ?

Results

Experiments with scripted + improvised recordings

Model (Input features)	Weighted Accuracy, WA	Unweighted Accuracy, UA
Model - 1 (acoustic)	**70.82**	55.53
Model - 2 (linguistic)	62.03	51.96
Model - 3 (fusion)	65.05	58.39
Model - 3 (downsampling + ensembling)	68.11	**63.15**

<u>Observation</u> :

- Improvement using Acoustic + Linguistic features ✓
- Improvement using Acoustic features or only Linguistic features ✗

Results

Experiments with scripted + improvised recordings

Model (Input features)	Weighted Accuracy, WA	Unweighted Accuracy, UA
Model - 1 (acoustic)	**70.82**	55.53
Model - 2 (linguistic)	62.03	51.96
Model - 3 (fusion)	65.05	58.39
Model - 3 (downsampling + ensembling)	68.11	**63.15**

Observation :

- Improvement using Acoustic + Linguistic features ✓
- Improvement using Acoustic features or only Linguistic features ✗

Reasoning :

- Class imbalance in the combined scenario plays important role
  - Model -1 achieves best WA but very low UA
- Fusion of linguistic information + acoustic features -> + 2.89% in UA

Results

Experiments with scripted + improvised recordings

Model (Input features)	Weighted Accuracy, WA	Unweighted Accuracy, UA
Model - 1 (acoustic)	**70.82**	55.53
Model - 2 (linguistic)	62.03	51.96
Model - 3 (fusion)	65.05	58.39
Model - 3 (downsampling + ensembling)	68.11	**63.15**

Observation :

- Improvement using Acoustic + Linguistic features ✓
- Improvement using Acoustic features or only Linguistic features ✗

Reasoning :

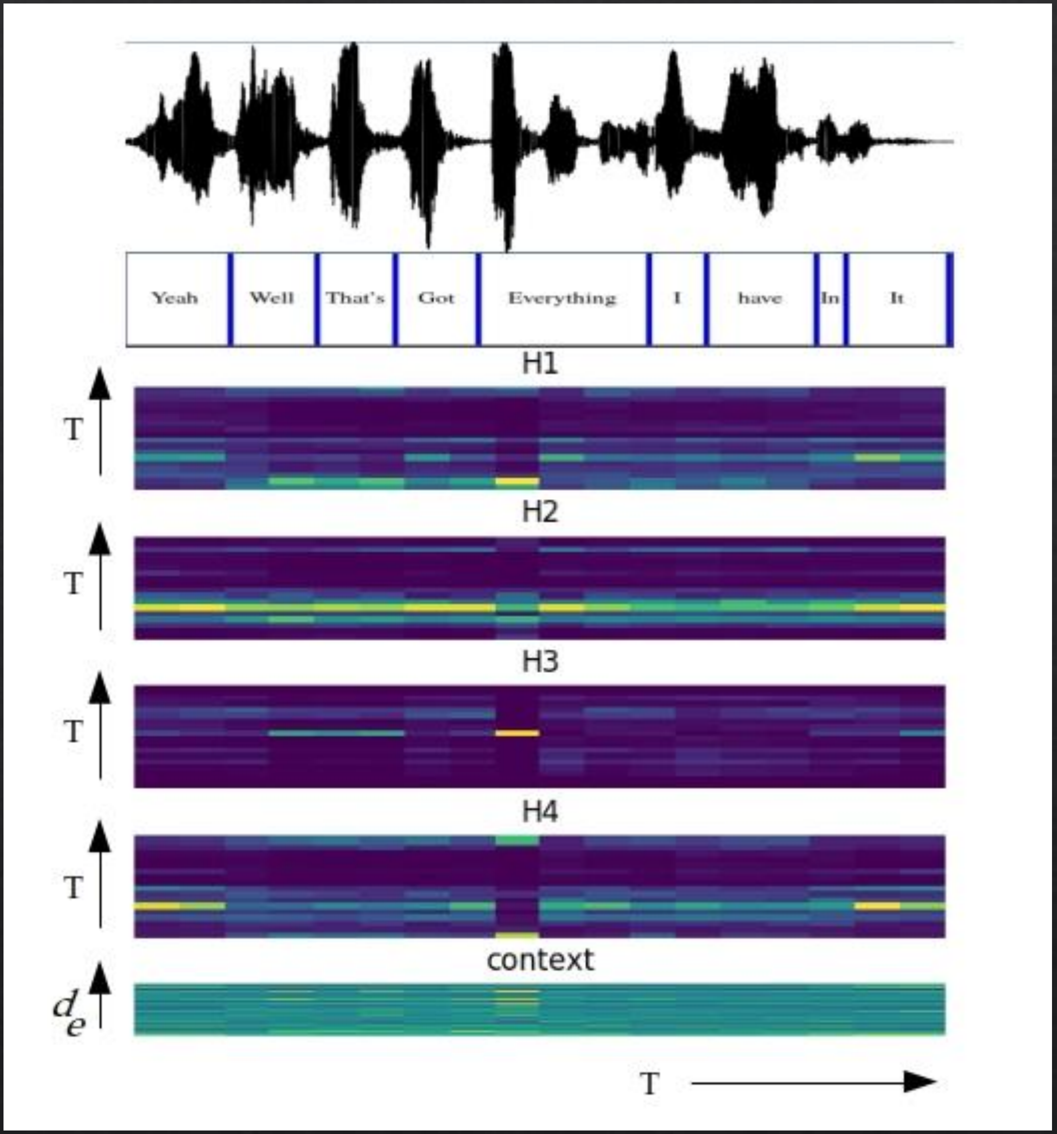
- Class imbalance in the combined scenario plays important role
  - Model -1 achieves best WA but very low UA
- Fusion of linguistic information + acoustic features -> + 2.89% in UA

But, is the self-attention module actually helping?

Discussion

$$A_i = softmax\left(\frac{\boxed{Q_i K_i^T}}{\sqrt{d_K}}\right) V_i$$

- Model learns the acoustically significant frames and weighs them heavily during the formation of context.
- Strong emphasis around the word “everything” makes it almost distinctive as anger emotion.
- Not all heads contribute equally, most important and confident heads play a consistent role.



Attention weights (a T X T matrix) for each attention head.
T : timesteps, True emotion : *anger*

Conclusion

- Proposed an End-to-End model for an improved SER system using self attention mechanism.

Conclusion

- Proposed an End-to-End model for an improved SER system using self attention mechanism.

- Less correlation of linguistic cues with the emotion than its acoustic counterpart in the improvised recordings.

Conclusion

- Proposed an End-to-End model for an improved SER system using self attention mechanism.
- Less correlation of linguistic cues with the emotion than its acoustic counterpart in the improvised recordings.
- Combination of linguistic and acoustic features gives an improvement of
  - 6.29% for only scripted
  - 2.86% for combined scenario

  indicating usefulness of our approach.

Thank You !

- **Swapnil Bhosale**
- **Rupayan Chakraborty**
- **Sunil Kumar Kopparapu**

TCS Research & Innovation Mumbai, India

(bhosale.swapnil2, rupayan.chakraborty,
sunilkumar.kopparapu)@tcs.com